A Case Study on the Importance of Named Entities in a Machine Translation Pipeline for Customer Support Content

Miguel Menezes[1,2] **Vera Cabarrão**[3] **Pedro Mota**[3]
Helena Moniz[1,2] **Alon Lavie**[3]

[1] Universidade de Lisboa, Lisboa, Portugal

[2] INESC-ID, Lisboa, Portugal

[3] Unbabel, Lisboa, Portugal

{lmenezes, helena.moniz}@campus.ul.pt

{vera.cabarrao, pedro.mota, alon.lavie}@unbabel.com

Abstract

This paper describes research developed at Unbabel, a Portugal-based translation technology company, that combines MT with human post-edition and focuses mainly on customer service content. We aim to contribute to furthering translation quality and good-practices by exposing the importance of having a **continuously-in-development** robust Named Entity Recognition system that, among other advantages, supports General Data Protection Regulation (GDPR) compliance. Moreover, we have tested semi-automatic strategies that support and enhance the creation of Named Entities gold standards to allow a more seamless implementation of Multilingual Named Entities Recognition Systems. The project described in this paper is the result of a shared work between Unbabel´s linguists and Unbabel´s AI engineering team, matured over a year. The project should also be taken as a statement of multidisciplinarity, proving and validating the much-needed articulation between the different scientific fields that compose and characterize the area of Natural Language Processing (NLP).

1 Introduction

Customer support professionals deal with multiple issues and problems arising from human-interaction, from answering questions or responding to customer complaints, to processing orders and returns, as well as sharing information and services. They are, in a sense, a direct line between customers and service providers, so they must be efficient, fast, and overall understandable, all while working remotely. Unbabel enhances customer support abilities through the combination of a Machine Translation (MT) layer, coupled with human post-edition, allowing to combine the speed and scale of MT with the quality of human editing.

To that end, we focus on Named Entity Recognition processes that compose a vital part of the automatic translation pipeline, since they promote an increase in translation quality, and ensure 2018 data protection regulation compliance. To promote high MT performances, a Named Entity Recognition System (NER) was applied, enabling the identification of NEs in context, e.g., prediction of NEs according to its surroundings, while simultaneously categorizing the NE. The identified NEs are then automatically blocked for translation or automatically annotated as NE of interest for further processes such as localization. This step ensures a decrease in MT "hallucinations" (inadequate translations) (Lee et al., 2018), since NEs are often responsible for these severe MT mistranslations, which

© 2022 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

negatively impacts the overall translation quality, considered mostly as critical errors in terms of severity. There is a second step associated with the NE pipeline, the anonymization process. The anonymization guarantees that all the NEs corresponding to personal identifiable information (PII) are either replaced by an adequate placeholder, for example *Email*; *Phone Number*; *Reference Number;* or replaced by a semantic equivalent (Mota et al., 2022) in case the NE is a person's name. In the latter, the real name is replaced with a fictitious name that agrees in gender with the original one. This step has a four-fold goal: i) ensures customer sensitive data protection and prevents MT learning with PII information; ii) prevents MT mistranslation; iii) ensures gender agreement (specifically in the case of the replacement of names for semantic equivalents), and iv) guarantees document readability, which is particularly relevant for post-editors. In short, the application of NER is fundamental for enhancing translation quality and preventing personal data breaches, which can lead to fines for non-compliance cases.

Despite the aforementioned importance that NEs represent within a MT pipeline, their definition seems to be somehow elusive. The fact that there is not a unique definition of what constitutes a NE in the literature can be directly associated with the fact that they are structures with the needed plasticity and adaptability to be applied to different tasks. At the MUC, (Chinchor et al., 1997), named entities were defined as "unique identifiers"; in 2003 CoNNL shared task: Language-Independent Named Entity Recognition, they were described as "phrases that contain the names of persons, *organizations and locations."* (Sang and De Meulder, 2003), and for Nouvel et al. (2016) they are "*textual units corresponding to predefined semantic categories*". Despite the different definitions, they all seem to agree that a named entity functions as a referent (Jurafsky et al., 2020: 1); a linguistic object carrying relevant information in a document, needed, according to Nouvel et al. (2016: 10), to allow the computer system to "understand" documents.

Considering the importance of such structures within a document, we investigate an alternative approach to semi automatically generate training data (still requires manually annotation source language) for Named Entity Recognition (NER) models from parallel corpus (Agerri et al. 2018: 3533). This is important for the use case where we want to expand NER language coverage within a particular domain. The goal is to only require NE annotated data on the source side and automatically determine the correspondence in the translation. This avoids the time-consuming and high-priced human annotations necessary to train NER for a new language.

To achieve our goals, we benchmark different alignment models, and use their output to project NEs annotations from source to target text. We will show their impact in the English–German and English–Brazilian Portuguese language pairs as well as in the domains of tourism and technology.

2 Related Work

In the last few years, machine learning systems have been predominantly used to achieve state-of-the-art NER results and much has been developed since the early Message Understanding Conferences (MUC) initiatives. A continuous flow of proceeding works in the field, both in the industry and in a more academic environment, has yielded significant changes that go from new, high performance computational technologies related to the NER subtask itself, to new different applications and goals. These frameworks have been developed to accommodate particular objectives for particular domains, such as in the case of the healthcare industry (Tarcar et al., 2019), where NER models were used, for example, to extract structure information from unstructured Electronic Health Records (EHR).

Despite all technological advances, commonly used frameworks still heavily rely on human intervention to provide modeling features or heuristics to solve downstream NLP tasks. While solutions have been proposed to overcome the need for these handcrafted features (Santos and Guimarães, 2015: 1), the need of labeled data is still an obstacle. In cross-lingual applications, this problem is further aggravated with the cardinality of the number of necessary language pairs. When expanding NER language coverage, this problem can be tackled using named entities word alignment within parallel corpora. This information allows the transfer of NE annotations from a source sentence and its translation (Eskin et al., 2019). Recent work has shown impressive results with the application of new deep learning models, e.g., Transformers, based on an encoder/decoder architecture, mapping sentences to vectors, which result in a representation of the input sequence of words in the source language

(Vaswani et al., 2017). This has boosted the quality of NER and word alignment models.

Akbik et al. (2018) propose *contextual string embeddings* for the NER. The embeddings are pre-trained on large unlabeled corpora without any explicit notion of words and thus, fundamentally, model words as sequences of characters, *contextualized* by their surrounding text. Therefore, the same word will have different embeddings depending on its contextual use. This allows the embeddings to properly represent polysemic words, language specific prefixes and suffixes, and handle misspelled words. The approach achieved state-of-the-art results in the *CoNLL 2003* NER shared task.

Wang et al. (2019) propose the use of the M-BERT, Multilingual Bidirectional Encoder Representations from Transformers, for cross-lingual transfer without the need of a dedicated cross-lingual training objective and with no aligned data. Experiments were carried out in three different languages (Spanish, Hindi, and Russian) and showed that M-BERT generalizes well across languages for a variety of downstream tasks (Wu and Dredze, 2019), like NER and Part of Speech (POS) tagging. Extending this research line, mLUKE and ERICA enhance M-BERT with Named Entity capabilities, further improving the state-of-the-art in several NLP downstream tasks.

Eskin et al. (2019) propose a neural model for word alignment, integrated into a Transformer-based machine translation model for English–Chinese and English–Arabic. The model can be used to generate cross-lingual NE datasets via alignment projection of token-level annotations in a high-resource language to a low-resource language.

Modrzejewski et al. (2020) explores an approach to improve translation quality by conveying NE information through source factors in a machine translation model. The method showed an increase of 1% in the BLEU score, when using the WMT2019 standard test, and an increase of 12% when compared with a strong baseline for NE translation.

As stated above, several NER models have been proposed, some with the main goal of allowing off-the-shelf usage, such as Stanza, Google Cloud Natural Language, and Spacy. In all systems, a wide variety of NEs are taken into account, that range from Address; Date-Time; E-mail, Payment/Credit-Cards in case of Google, or Location; Facilities; Law; Language, *inter alia,* in case of Spacy. Nevertheless, performing NER in a specific domain remains a challenge. In our case, we target the customer-support domain, where the previous tools underperform or lack necessary NE types. We resort to training custom models with in domain data. Scaling this approach to many different languages is expensive due to the cost of obtaining labeled data. By using a word alignment-based approach (Chung, 2007: v) to project existing NE annotations to a new language in parallel corpora we can address this issue.

3 Dataset Annotation

To validate the word alignment-based NE projection, we manually annotated two datasets: Tourism-Dataset, and Technology-Dataset. For the Tourism-Dataset, we used parallel data (bitext) in EN (source) and in DE. The datasets, comprising 2500 sentences each, were annotated by two linguists, one responsible for the EN data set annotation, whilst a second one was responsible for the DE version. For DE two different translations were annotated, one from machine translated only (MT), and the other with an extra post-edition layer (PE). The Technology-Dataset consists of 360 post-edited sentences for the EN–PT/BR language pair and was fully annotated by one of the previous linguists.

All datasets went to a preprocessing stage, where the data sets were divided into sentences, allowing the annotation to be made sentence by sentence using Prodigy[2], an annotation platform. Both annotators used Unbabel´s internal NE annotation guidelines. The annotators also had access to online information, namely dictionaries, maps, and other relevant sources of information that could facilitate the task.

3.1 Named Entities Typologies

For the Named Entity Recognition task, it is important i) to define which NEs are relevant for the job and ii) how to annotate them. This process requires the creation of a NE typology, "a descriptive formalization of the selected categories and their scope" (Nouvel et al., 2016: 48), that usually comes in the form of annotation guidelines. This project uses the current generic NEs typology created by Unbabel, that follows the universal Named Entity categories triad: Enamex,

[2] https://prodi.gy

Numex and Timex (Table 1 shows the complete NEs categories tag set applied in this study).

3.2 Inter-annotator agreement

Given that the linguists worked separately in the Tourism-Dataset, we carried out an inter-annotator agreement study to determine if the NE typology was similar in the corresponding EN/DE language pair.

For the following analysis, we only considered a NE match within both gold standards whenever both annotators agreed in: i) the entity span, and ii) the category. The analysis performed allowed us to identify a high inter-annotator agreement, between the EN gold standard (source), and the two DE datasets (target): 90% for the MT and 91% MT with PE.

Named Entities Categories	Named Entities Inter-Annotator Agreement Results		
	EN GS	DE MT GS	DE PE GS
Organization	183	161	167
Currencies	284	276	278
Percentages	9	9	9
Refnumber	64	52	53
Names	45	43	43
Dates	106	102	102
Address	26	22	23
E-mail	12	12	12
Phone Number	15	15	15
Time	26	21	21
URL	18	17	17
City	56	39	39
Country	3	3	3
Products and Services (PRS)	13	4	4
Credit Card	1	1	1
Password	1	1	1
Username	1	1	1
Number Code	1	0	0
Total	865	781	789

Table 1: NEs inter-annotator agreement in absolute values.

By observing the EN gold standard, we were able to account for 865 named entities identified by annotator one and 781 NEs identified by annotator two for DE MT gold standard, and 789 for the DE PE gold standard (Table 1). By pairing the number of identified NEs between the EN and DE gold standards, we determined that annotator two annotated less 9.72% NEs in the MT and less 8.72% NEs in the post-edited dataset than the total amount of NEs found in the EN gold standard, however, with very high inter-agreement in specific named entities, namely expressions that identify numbers (Numex NEs), such as:

1. Percentages: 100% agreement between EN and both DE gold standards.

2. Currencies: 97.1% agreement in MT and 97.8% in PE;

3. Phone numbers: 100% agreement.

Temporal expressions, Timex, e.g. Dates or Time, seem to follow the same pattern, amounting to a 96.22% agreement value in case of dates, and 80.76% for the category time, both in MT and PE. For Enamex entities, countries had 100% of inter-annotator agreement, and person names presented a value of 95%. There seems to be an intuitive understanding of these categories, corroborated by the lexical material in its surroundings, helping to assert such entities with fewer annotation doubts, as seen in the following examples taken from our datasets:

Ex.1

EN: "Dear Manuela Frieda Kalo"

DE: "Sehr geehrte(r) Manuela Frieda Kalo"

Greetings like in the above example, *Dear ...*, or in German *Sehr geehrte(r)...,* hint that the following word is a named entity, specifically a name, being relevant both for the human-annotation process and for the MT system learning process.

Based on the annotation agreement values for the above-mentioned categories, we conclude that all these NEs gather consensus; they tend to be context-independent and, hence, straightforward to annotate. In these cases, there are few doubts as to which tags to choose. On the other hand, the NEs labeled as *Products and Services* (PRS)

present the lowest inter-annotation agreement score, 30%. Many of the named entities labeled as PRS in the EN gold standard were tagged as *Organizations* (ORG) both in MT and PE DE gold standards, thus being considered mismatching NEs. Moreover, for these categories, the same NE can assume both categories in different sentences, thus denoting ambiguous characteristics. In these cases, interpreting the entire sentence, or the words in a NE vicinity can be the key to determine its role and classification. However, this approach might not always be so linear or straightforward, as shown in the following examples:

Ex.2
EN: "Kindly make sure that one of the accepted cards like [Union pay credit card]$_{\textbf{Organization}}$ is saved in your [HolidayConsultee]$_{\textbf{Organization}}$ account."

DE: "Bitte stellen Sie sicher, dass eine der akzeptierten Karten [Union Pay Kredit-, die HolidayConsultee --Karte]$_{\textbf{Products and Services}}$ in Ihrem-Konto gespeichert ist."

In the cases above, every single NE was identified as an ORG in the EN gold standard, while in the DE gold standard, they were tagged as PRS. The annotation differences reside on the fact that in the EN gold standard, the named entity was taken by the annotator one as an entity that provides a service, whereas in the DE gold standard, the annotator two interpreted the named entity as a service itself.

Overall, we can define the inter-annotator agreement for this task as substantially high, nevertheless, we must accept the fact that for some categories, like PRS, and ORG and even *Locations* (LOC), the annotation task is not fully consensual, leading to inter-annotator mismatches.

4 Named Entity Projection

To understand the impact of using an alignment approach in building a multilingual NER system, we tested four state-of-the-art aligners: FastAlign[3], the current aligner used by Unbabel; eflomal[4]; SimAlign[5], and AwesomeAlign[6]. Each aligner had available different sets of configurations that, when combined, amounted to a total of 53 different alignment possibilities for each NE category. The different configuration for aligners ranged from:

· Heuristics, allowing different alignment directions: from source to target and vice versa, with the goal (Mota et al., 2022);

Training data that range from more generic data to client data or mixed data (both generic and client data); or

· Pre-trained models for cross-lingual understanding.

Using the output word alignments, NE identified in the source sentence were projected in the target based on a min-max algorithm. This means that we consider the target entity span to range the lowest to highest word alignments.

Model ranking for NE projection task results were presented for assessment using an online software, developed by Unbabel´s AI team, that showed all alignment results for the four aligners used, together with their configurations. The alignment results were displayed from best (number 0) to worst alignment result (number 53). Moreover, the developed interface also allowed us to compare two models (Figure 1), giving a panorama over the alignment quality for each category (Figure 2).

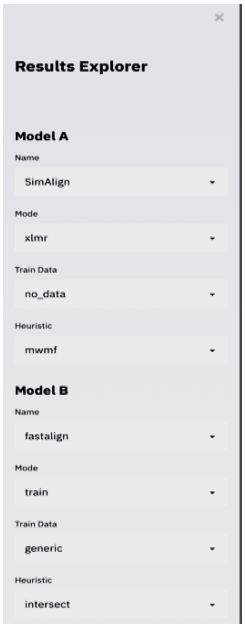
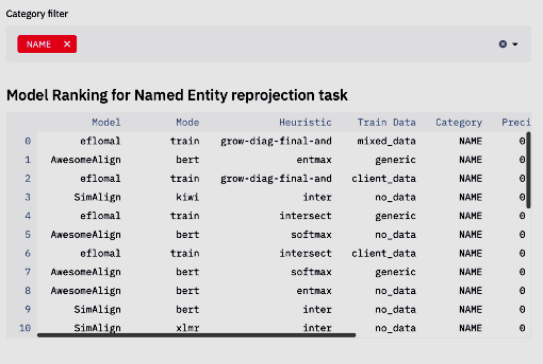


Figure 1: Aligners model comparison, giving the ability to choose between available configurations.

[3] https://github.com/dgel/fastalign
[4] github.com/robertostling/eflomal
[5] github.com/cisnlp/simalign
[6] github.com/neulab/awesome-align



	Model	Mode	Heuristic	Train Data	Category	Preci
0	eflomal	train	grow-diag-final-and	mixed_data	NAME	0
1	AwesomeAlign	bert	entmax	generic	NAME	0
2	eflomal	train	grow-diag-final-and	client_data	NAME	0
3	SimAlign	kiwi	inter	no_data	NAME	0
4	eflomal	train	intersect	generic	NAME	0
5	AwesomeAlign	bert	softmax	no_data	NAME	0
6	eflomal	train	intersect	client_data	NAME	0
7	AwesomeAlign	bert	softmax	generic	NAME	0
8	AwesomeAlign	bert	entmax	no_data	NAME	0
9	SimAlign	bert	inter	no_data	NAME	0
10	SimAlign	xlmr	inter	no_data	NAME	0

Figure 2: Best alignments scoring for the Name category, considering the different model´s configurations (Mode; Heuristic; Train Data)

With access to the information displayed by the above-mentioned interface, we were able to understand the differences in alignments that generated NEs spans between the EN source dataset and its DE counterpart. Moreover, we were also able to compare the DE dataset with and without an extra post-edition layer, as to determine if such a task does interfere positively or negatively in the NE projection results. Also, we were able to evaluate the aligner settings that showed better performance within the 53 possible combinations and benchmark the current aligner used by Unbabel. The NE projection task was evaluated using a classification setting with the following standard performance metrics: Precision, Recall and F_1 (Makhoul et al., 1999), in order to have a more fine-grained performance perspective of the applied model results:

The precision value is defined as the number of positive NE predictions (true positives) divided by the sum of true positives and false positives. This formula is used to understand the classifier exactness. The question that the concept of precision answers is, of all the NEs retrieved by the NE projection algorithm, how many were actually correct. Lower values of precision indicate a higher number of false positives.

The recall value is defined as the ratio of correctly predicted true positive NEs, divided by the sum of true positives and false negatives. The question recall answers is, of all the NEs in the test dataset, how many were retrieved correctly by the NE projection algorithm.

The F-value, also known as F_1, is defined as the harmonic mean of the precision and the recall, being appropriate to identify the desired average rate.

5 Experimental Results

Our study yields very promising results, showing the devised approach to be trustworthy for building multilingual gold standards for NER training when the correct alignment system coupled with specific correct configurations is implemented.

5.1 Tourism Dataset

This section provides the NE projection results obtained for the Tourism-Dataset. Based on the F_1 results obtained for each NE category, we are able to determine the best performing aligner. The overall results can be found in Table 2.

	SimAlign	FastAlign	AwesomeAlign	eflomal
N	6	5	3	3

Table 2: Number of categories for which each alignment system achieved the best alignment results.

Based on these results analysis, we were able to ascertain that SimAlign proved to be the best alignment model for six categories: *Organization, Currency, City, Time, Products and Services and Dates,* generating the most trustworthy alignments using the XLM-R pre-trained model and the intersect symmetrization heuristic.

FastAlign was ranked as second-best aligner, obtaining top alignments for the following categories: *Country, Credit card, Address, Percentages, Username*. The remaining six categories' first place alignments were divided between the remaining two aligners, eflomal and AwesomeAlign, which led us to immediately discard them as top aligners. The alignment results analysis also led us to conclude that SimAlign behaves in a very consistent manner, obtaining very high F_1 scores overall.

A more in-depth analysis for the *Currency* category can be found in Tables 3 and 4. The first table displays the top five best overall alignment results. The second one, dedicated exclusively to the aligner currently used by Unbabel, FastAlign, displays the top five best alignment configurations. Based on these results, we can state that, for the *Currency* category, SimAlign outperformed the remaining aligners, producing the five best alignment results overall. On the other hand, FastAlign only ranked in 17th place (and onwards) for NE projection, resulting in an

alignment quality difference between both aligners of 0.076%.

Model	Mode	Heuristic	Train data	Categ.	Precision	Recall	F_1	Time
SimAlign	**Bert**	**Inter**	**No data**	**CRR**	**0.981**	**0.975**	**0.976**	**0.0205**
SimAlign	kiwi	Inter	No data	CRR	0.981	0.974	**0.974**	0.0284
SimAlign	kiwi	itermax	No data	CRR	0.976	0.978	**0.974**	0.318
SimAlign	xlmr	mwmf	No data	CRR	0.976	0.977	**0.973**	0.4719
SimAlign	kiwi	mwmf	No data	CRR	0.976	0.977	**0.973**	0.3695

Table 3: Top five alignment results for the *Currency* NE.

Model	Mode	Heuristic	Train data	Categ	Precision	Recall	F_1	Time
FatsAlign 17th	production	Grow diag final	No data	CRR	0.934	0.894	0.899	0.0007
FatsAlign 18 th	production	intersect	No data	CRR	0.973	0.853	0.889	0.0007
FastAlign 19 th	train	Grow diag final	Mixed data	CRR	0.914	0.883	0.883	0.0005
FastAlign 20 th	train	Grow diag final	generic	CRR	0.906	0.881	0.878	0.0005
FastAlign 21st	train	intersect	Mixed data	CRR	0.975	0.824	0.866	0.0005

Table 4: Top five best alignment results for FastAlign for the *Currency* NE.

5.2 Technology Dataset

This section provides the NE projection results obtained for the Technology-Dataset. The analysis is displayed for each category within the parallel *corpus.*

For the category *Name,* SimAlign and AwesomeAlign reached constant F_1 values of 1, regardless of the configurations applied. On the other hand, 39.29% of the alignments carried out by FastAlign and Eflomal were deemed having F_1 value of under 1.

For *Currency*, the results for SimAlign and AwesomeAlign followed the same pattern, while FastAlign and eflomal never reached a F_1 value over 0.75.

For the category *Organizations,* once again AwesomeAlign and eflomal reached constant values of 1. SimAlign and FastAlign results ranged between 0.91 to 1. The configuration responsible to SimAlign underachievement reads as follow:

- Mode: BERT
- Heuristic: Itermax

For the category *Email*, all alignment-based NE projection results were deemed as having F_1 scores of 1, except for the ones performed by FastAlign with 50% of the all alignments with a F_1 of 0.

Regarding the category *URL,* all models reached F values of 1, except FastAlign with constant values under 0.66.

As for *Products and Services,* the overall F value results ranged between 0.58 and 0.97. Nevertheless, we were still able to ascertain solid F_1 scores of 0.97 for AwesomeAlign and SimAlign.

For the category *Reference Number* AwesomeAlign, SimAlign and eflomal alignments reached constant values of 1. FastAlign underperformed reaching a top value of 0.75.

The previous results show that AwesomeAlign produced the best NE projections, followed by SimAlign that only for the category *ORG* did not show an F_1 of 1. AwesomeAlign configurations produced alignments with F_1 results of 1, similarly to SimAlign, (excluding the category *PRS,* as previously mentioned), suggesting that the task was trivial to solve. Nevertheless, it is important to highlight that the dataset used for alignment only comprised 360 sentences, with a very small amount of NEs per category. Moreover, most of the NEs had a similar form in both source and target, making the projection task easier. The lack of enough NEs representing a category can explain the F_1 obtained by AwesomeAlign and SimAlign, independently of the particular configuration.

With regards to FastAlign, it still underperforms in comparison with the other aligners, being for some categories the aligner that presented the worst alignment results. We hypothesize that the underperformance of FastAlign is related to its difficulty in dealing with rare words, which typically are instances of NEs. The pre-trained model-based approaches are more robust when facing this issue since they operate at the subword level and are exposed to much larger datasets during training.

6 Conclusions and Future-work

With this work, we focused on giving a general overview on the pivotal importance of NEs from a linguistic and historical perspective, highlighting its relevance within an automatic-translation scenario. Moreover, we were able to test four different aligners for the creation of semi-automatic multilingual gold standards through NE projection in parallel *corpora.* With the research results concerning the creation of multilingual gold standards, we were able to replace the aligner used in production, Fastalign, by SimAlign. By doing so, we ensure a reliable integration of this cross-lingual technique for the creation of multilingual **NER gold standards** for multiple language pairs and applicable to a myriad of different domains. The manual-annotation tasks performed along the experiments also allowed us to highlight the fact that particular NEs can play ambiguous roles and can be responsible for inter-annotator mismatches, thus needing special attention.

Also, we see future possibilities of using the NER system to leverage Unbabel´s Translation Memories. The identification of NEs followed by their replacement with corresponding placeholders will lead to an increase in the number of Translation Memories matches, which promotes more accurate end-translation results, while lessening, simultaneously, the need for human post-edition.

Finally, a note still on the contribution of our work to the anonymization module in the pipeline. The NE work conducted ultimately reflects improvements on the anonymization module, crucial to any company compliant with Responsible AI Principles. The fundamentals and approaches developed within our project regarding the identification and anonymization of Personal Identifiable Information have already been implemented by the MAIA Project (Multilingual AI Agent Assistants), thus enabling information processing and sharing in a safe manner. As such, we will continue our work concerning the NER task, with a particular focus on the anonymization step.

Acknowledgements

This work was supported by national funds in Portugal through Fundação para Ciência e a Tecnologia (FCT), with reference UIDB/50021/2020 and through FCT and Agência Nacional de Inovação with the Project Multilingual AI Agents Assistants (MAIA), contracted number 045909.

References

Agerri, Rodrigo, Yi-Ling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 3529-3533.

Akbik, Alan, Duncan Blythe and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In Proceedings of the 27th international conference on computational linguistics (pp. 1638-1649).

Chinchor, Nancy and Patricia Robinson. 1997. Message Understanding Conference-7 named entity task definition. In Proceedings of the Seventh Conference on Message Understanding (Vol. 29, pp. 1-21).

Chung, Yi-Ling. (2017). Automatic generation of named entity taggers leveraging parallel corpora. Stanford University.

Data Protection Act, 2018. Data Protection Act 2018. [online] GOV. U.K. Available at: <https://www.gov.uk/government/collections/data-protection-act-2018>

Finkel, J. Rose, Trond Grenager and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05) (pp. 363-370).

Joseph, Cris (2019, March 1) What are the Benefits of delivering excellent customer service?. Chron. Retrieved January 18, 2020, from 2019 https://smallbusiness.chron.com/benefits-delivering-excellent-customer-service-2086.html

Jurafsky, Dan and James H. Martin. 2018. Speech and Language Processing. *Chapter 8: Sequence Labelling for Parts of Speech and Named Entities* (draft of December, 30, 2020).

Lee, Katherine, Orhan Firat Ashish Agarwal, Clara Fannjiang and David Sussillo. 2018. Hallucinations in Neural Machine Translation. Google AI. Retrieved January 18, 2020, from Openreview.net, Available at https://openreview.net/forum?id=SkxJ-309FQ

Makhoul, John, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measure for information extraction. In Proc. of the DARPA Broadcast News Workshop, Herndon, VA.

Modrzejewski, Maciej, Miriam Exe. Bianka, Buschbeck, Thanh-Le Ha and Alexander Waibel.

2020. Incorporating external annotation to improve named entity translation in NMT. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (pp. 45-51).

Mota, Pedro, Vera Cabarrão and Eduardo Farah. 2022. Fast-Paced Improvements to Named Entity Handling for Neural Machine Translation. In Proceedings of EAMT.

NER Annotation Guidelines. (2020). Unbabel´s Internal Company Document.

Nouvel, Damien, Maud Ehrmann and Sophie Rosset. 2016. Named entities for computational linguistics. ISTE.

Qin, Yujia., Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun and Jie Zhou. 2020. Erica: Improving entity and relation understanding for pre-trained language models via contrastive learning. arXiv preprint arXiv:2012.15022.

Ri, Ryokan, Ikuya Yamada and Yoshimasa Tsuruoka. 2021. mLUKE: The Power of Entity Representations in Multilingual Pretrained Language Models. arXiv preprint arXiv:2110.08151.

Sang, Erik and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. arXiv preprint cs/0306050.

Santos, N. D. Cicero and Victor Guimarães. 2015. Boosting named entity recognition with neural character embeddings. arXiv preprint arXiv:1505.05008.

Stengel-Eskin, Elias, Tzu-Ray Su, Matt Post and Benjamin Van Durme. 2019. A discriminative neural model for cross-lingual word alignment. arXiv preprint arXiv:1909.00444.

Tarcar, K. Amogh, Aashis Tiwari, Vineet Naique Dhaimodker, Penjo Rebelo, Rahul Desai and Dattaraj Rao. 2019. Healthcare NER models using language model pretraining. arXiv preprint arXiv: 1910.11241.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones., Aidan N. Gomez, Lukasz kaiser and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Wang, Zihan, Stephen Mayhew and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. arXiv preprint arXiv:1912.07840.

Wu, Shijie and Mark Dredze. 2019. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT. EMNLP arXiv:1904.09077.